Goal-oriented quantization: Analysis, design, and application to resource allocation

Hang Zou, Chao Zhang, Samson Lasaulce, Lucas Saludjian, Vincent Poor

Goal-Oriented Quantization: Analysis, Design, and Application to Resource Allocation

Hang Zou, Chao Zhang, Samson Lasaulce, Lucas Saludjian, and Vincent Poor

***Abstract*—In this paper, the situation in which a receiver has to execute a task from a quantized version of the information source of interest is considered. The task is modeled by the minimization problem of a general goal function $f(x; g)$ for which the decision x has to be taken from a quantized version of the parameters g. This problem is relevant in many applications e.g., for radio resource allocation (RA), high spectral efficiency communications, controlled systems, or data clustering in the smart grid. By resorting to high resolution (HR) analysis, it is shown how to design a quantizer that minimizes the gap between the minimum of f (which would be reached by knowing g perfectly) and what is effectively reached with a quantized g. The conducted formal analysis both provides quantization strategies in the HR regime and insights for the general regime and allows a practical algorithm to be designed. The analysis also allows one to provide some elements to the new and fundamental problem of the relationship between the goal function regularity properties and the hardness to quantize its parameters. The derived results are discussed and supported by a rich numerical performance analysis in which known RA goal functions are studied and allows one to exhibit very significant improvements by tailoring the quantization operation to the final task.**

***Index Terms*—Goal-oriented communications, semantic communications, high resolution quantization, clustering, Bennett's integral, Gersho's conjecture.**

I. Introduction

Since the pioneering and fundamental works of Shannon [1], the dominant paradigm for designing a communication system is that communications must satisfy quality requirements. Typically, the bit error rate, the packet error rate, the outage probability, or the distortion level must be minimized. It turns out that the conventional paradigm consisting in pursuing communication reliability or possibly security may not be suited to scenarios such as systems where communications occur in order for a given task to be executed. For instance, transmitting an image of 1 Mbyte to a receiver that only needs to decide about the absence/presence of a given object in the image might be very inefficient. In this example, the receiver only needs one bit of information and this bit could have been directly sent by the transmitter and make the use of the communication and computation resources much more efficient. This simple example shows the potential of making a communication task- or goal-oriented (GO).

In this paper, the focus is on the problem of signal compression when the compressed signal is used for a given task which is known. More precisely, we focus on the signal quantization problem, which is often a key element of a signal transmitter. Introducing and developing a goal-oriented quantization (GOQ) approach is very relevant for many applications. We will mention three of them. First, it appears in controlled networks that are built on a communication network. A simple example is given by modern power systems such as the smart grid. A data measurement system such as a smart meter may have to quantize or cluster the measured series for complexity or privacy reasons [2]. It is essential that the quantization or clustering operation does not impact too much the quality of the decision (e.g., a power consumption scheduling strategy) taken e.g., by an aggregator. Second, GOQ is fully relevant for wireless RA problems. For instance, if a wireless transmitter receives some quantized information from the receivers/sensors through a limited-rate feedback channel [3]–[7]. Third, for future wireless communication systems such as 6G systems [8]–[11], GOQ and more generally GO data compression constitutes a very powerful degree of freedom of increasing final spectral efficiency since only the minimum number of bits to execute the task is transmitted through the radio channel.

The conventional quantization approach [12] is to minimize some distortion measure between the original signal and its representation, regardless of the system task. In the literature, there exist works on the problem of adapting the quantizer to the objective. For instance, in the wireless literature, the problem of quantizing channel state information (CSI) for the feedback channel has been well studied (see e.g., [13] for a typical example). The practical relevance of low-rate scalar quantizers to transmit high dimensional signals has been defended for MIMO systems in [14] [15] [16]. By combining the system task with the quantization process, [17] [18] investigated the influence of scalar quantization on specific tasks and characterized the limiting performance in the case of recovering a lower dimensional linear transformation of the analog signal and reconstruction of quadratic function of received signals. Deep-learning-based quantizers have also be considered in [19]–[22] to adapt to the task by training neural networks. The main point to be noticed is that for all existing works either the impact of quantization on a given performance metric is studied or a very specific performance metric is considered (the Shannon transmission rate being by far the most popular metric) and the proposed quantizer design is often an ad hoc scheme. In contrast with this line of research works, we introduce a general framework for GOQ illustrated in Fig. 1. The task or goal of the receiver is chosen to be modeled by a generic optimization problem (OP)

which contains both decision variables and parameters. One fundamental point of the conducted analysis is that both for the performance analysis and the design, the goal function is a generic function $f(x;g)$, x being the decision with dimension d to be made based on a quantized version of the function parameters g with dimension p. This setting allows us to derive analytical results and acquire completely new insights into how to adapt a quantizer to the goal, these insights relying in part on the high resolution (HR) regime analysis [23]–[25].

To be sufficiently complete concerning the technical background associated with the present contributions, we also would like to clearly position our works w.r.t. recent works on semantic communications [26]–[38]. Semantics is employed here with its etymological meaning, that of significance. It can be seen as a measure of the usefulness/importance of messages with respect to the system task [26]. There have been several tutorials and surveys to discuss possible structures and architectures of this novel communication paradigm. By studying the semantic encoder and semantic noise, [27] proposed two models based on shared knowledge graph and semantic entropy, respectively. Reference [28] indicated that by properly recognizing and extracting the relevant information to the system task, the communication efficiency and reliability can be enhanced without using more bandwidth. In [26], it is explained how semantic information attributes of transmitted messages could be exploited, which entails a task-oriented unification of information generation, transmission, and reconstruction. By introducing intrinsic states and extrinsic observations, [29] uses indirect rate-distortion theory to characterize the reconstruction error of semantic information induced by lossy source coding schemes. Information bottleneck is also an approach to find the optimal tradeoff between compressing and reliability. Inspired by this approach, [30] proposed a relevant loss function whose relevance was supported in [31] and designed an end-to-end DeepSC network architecture, using Transformer as the semantic encoder and joint source-channel coding schemes to ensure the semantic information transmission. Similar models [32] [33] are extended to audio transmission and Internet-of-things (IoT) applications. Other learning tools have also been implemented to extract important attributes in semantic communications, such as reinforcement learning [34], curriculum learning [35], and distributed learning [36] [37]. Some additional information can also be used for the semantic encoder, such as contextual reasoning [38]. Compared to the quoted works, three main points have to be noticed. First, most works focus on the novel communication architecture or use learning tools to extract important features but the works are not supported by theoretical derivations. Second, we not only consider the transmission problem of the semantic information but also the influence of distorted information on the subsequent decision-making (DM) entity and the system task, namely, how the semantic information exchange will affect the system performance (effectiveness level). Third, we address a precise technical problem which is the quantization problem and assume a fully generic goal. The closest contributions to the present work have been produced by the authors through [39] [40] [41] [52]. To the best of the authors knowledge, the concept of GOQ has been introduced for the first time in [39] and applied in other contexts in [40] [41] [52]. In these references, mainly numerical results are provided and the focus is on a Lloyd-Max (LM)-type algorithm [54] [55]. In particular the formal HR analysis is not conducted and the fundamental role of the goal function is not investigated.

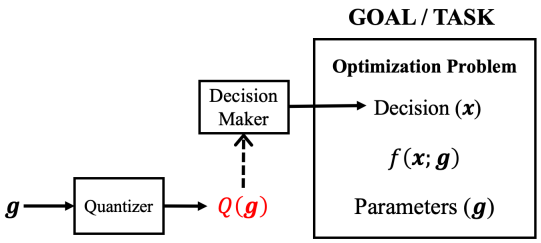


Fig. 1: Proposed definition for the goal-oriented quantization approach

This paper is structured as follows. In Sec. II, we define the performance metric of a GO quantizer. In Sec. III, the performance analysis of scalar GOQ is conducted in the HR regime and the impact of the goal function on the optimality loss (OL) is assessed through analytical arguments. In Sec. IV, we address the more challenging case of vector GOQ by providing an HR equivalent of the HR OL and a practical GOQ algorithm. In Sec. V, we show the potential benefit from using GOQ for important RA problems that are relevant for quantizing information in wireless, controlled, and power systems. Sec. VI concludes the paper.

II. Problem Formulation

Definition II.1. *Let $d \geqslant 1$ be an integer and $\mathcal{G}$ be a subset of $\mathbb{R}^d$. Let $M \geqslant 1$ be an integer. An $M-$quantizer $\mathcal{Q}_M$ is fully determined by a piecewise constant function $Q_M : \mathcal{G} \to \mathcal{G}$ that is defined by $Q_M(g) = z_m$ for all $z_m \in \mathcal{G}_m$ where: $m \in \{1, ..., M\}$, the sets $\mathcal{G}_1, ..., \mathcal{G}_M$ are called the quantization regions and define a partition of $\mathcal{G}$, and the points $z_1, ..., z_M$ are called the region representatives.*

Since M is a fixed number, from now on and for the sake of clarity, we will omit the subscript M from the quantization function and merely refer to it as Q. We will only make M appear for comparison purposes, mainly in the simulations. Also, when needed, we will also use the quantity $R = \log_2 M$ which represents the number of quantization bits per sample. Equipped with these notations, we can now define mathematically the GO approach we propose for quantization.

Definition II.2. *Let $\chi(g)$ be the decision function providing the minimum points for the goal function $f(x;g)$, whose decision variable is $x \in \mathbb{R}^p$ ($p \geqslant 1$ is an integer), g being fixed:*

$$\chi(g) \in \arg\min_{x \in \mathcal{X}} \quad f(x;g). \tag{1}$$

The optimality loss induced by quantization is defined by:

$$L\left(Q;f\right)=\alpha_f\int_{g\in\mathcal{G}}\left[f\left(\chi\left(Q\left(g\right)\right);g\right)-f\left(\chi\left(g\right);g\right)\right]\phi\left(g\right)\mathrm{d}g \tag{2}$$

where ϕ is the probability density function (p.d.f) of g and $\alpha_f>0$ is a scaling/normalizing factor which does not depend on Q.

Several comments concerning the OL definition are in order. Note that the conventional quantization approach can be obtained from the GOQ approach by observing that the second term of the OL functional $L(Q;f)$ (that is, a function of function) is independent of Q and by specializing f as $f(x;g)=\|x-g\|^2$, $\|.\|$ standing for the Euclidean norm. With the conventional approach, quantization aims at providing a version of g that resembles to g. However, under the GOQ approach, what matters is the quality of the end decision taken. The design of such a quantizer therefore depends on the mathematical properties of f and the underlying decision function χ, which constitutes a key difference w.r.t. the conventional approach. In this respect, studying analytically the relationship between the nature of f and the quantization performance is a nontrivial problem. For instance, for a fixed OL level, how do the functions requiring a small (resp. large) M (that is, a small -resp. large- amount of quantization resources) look like? The normalizing factor α_f is precisely introduced to conduct fair comparisons between different goal functions. From the OL definition, it can also be noticed that the knowledge of the p.d.f. of g is implicitly assumed. One may replace the statistical mean with an empirical mean version and rewrite the OL under a data-based form where the integral is replaced with a sum over the data samples obtained from a training set. Indeed, the knowledge of the input distribution ϕ is indeed convenient, especially for the analysis. However, for the design it is not required. This is why the proposed GO quantization algorithm is applied to the problem of data clustering, in which only a database is available. The case of a time-varying input distribution is not addressed here and would require to design an adaptive quantizer, which is left as a relevant extension of the present work. Also note that the set $\mathcal{X}$ and the function $\chi(g)$ are assumed to integrate the possible constraints on the decision x. At last, note that when the optimal decision function (ODF) $\chi(g)$ is not available, other decision functions that are suboptimal but easier to implement may be considered; this situation will be studied in the numerical analysis.

In what follows, the main focus is on the regime of large M, which is called the high resolution regime. This regime is not only very useful to conduct the analysis and make interpretations but also to provide neat approximants or expressions. These expressions are both exploited to obtain useful insights for the design of general quantizers and used in the proposed quantization algorithm. As it will be seen in the numerical performance analysis, the proposed algorithm performs remarkably well in the low resolution regime. Note that the direct minimization of the general form of the OL is an NP-hard problem since it is a mathematical generalization of the conventional quantization problem (see e.g., [20], [42]). Therefore, using approximants and suboptimal procedures is a classical approach in the area of quantization especially for vector quantization.

III. Scalar GOQ in the high resolution regime

In this section we assume that both the decision to be taken and the parameter to be quantized are scalar that is, $d=p=1$. For a wireless communication, this would occur for instance when a receiver has to report a scalar channel quality indicator (such as the SINR, the carrier/interference ratio, or the received signal power) to a transmitter and the transmitter tunes in turn its transmit power. Similarly, a real-time pricing system [43] in which an electrical power consumer reports its time-varying satisfaction parameter to an aggregator who chooses the price dynamically corresponds to the scalar case. Additionally, many systems, for complexity reasons, implement a set of independent scalar quantizers instead of a vector one. This is the case for example for some image compression standards such as JPEG or for MIMO communications with quantized CSI feedback [44]–[46]. In the general case, finding a quantizer amounts to finding both the regions $\mathcal{G}_1,...,\mathcal{G}_M$ (which are just intervals in the scalar case) and the representatives $z_1,...,z_M$. However, the calculation of regions and representatives can be simplified in the HR regime. One could use probabilistic density function to represent the density of quantization points, which allows us to approximate summations by integrals. To be precise, we assume the HR regime in the following sense [12]. For any point g, let us introduce the quantization step $\Delta(g)=\min_{1\leqslant m\leqslant M}|g-z_m|$. Then, let us introduce the (interval/representative) density function $\rho(g)$ which is defined as follows:

$$\rho(g)=\lim_{M\to+\infty}\frac{1}{M\Delta(g)}. \tag{3}$$

A. Optimal quantization interval density function

By construction, the number of quantization intervals or representatives in any interval $[a,b]$ can be approximated by $M\int_a^b\rho(g)\mathrm{d}g$. Therefore, the problem of finding a GOQ in the HR regime amounts to finding the density function that minimizes the OL that we will denote, with a small abuse of notation but for simplicity by $L(\rho;f)$. Remarkably, the expression of the optimal density in the HR regime can be obtained, at least by assuming the goal and decision functions to be sufficiently regular or smooth. This is the purpose of the next proposition.

Proposition III.1. *Let f be a fixed goal function. Assume f κ times differentiable and χ differentiable with*

$$\kappa=\min\left\{i\in\mathbb{N}:\ \forall g,\ \left.\frac{\partial^i f(x;g)}{\partial x^i}\right|_{x=\chi(g)}\neq 0\ \text{a.s.}\right\}. \tag{4}$$

In the HR regime the OL $L(\rho;f)$ is minimized by using the following quantization interval/representative density function:

$$\rho^{\star}(g) = C\left[\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa} f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\right]^{\frac{1}{\kappa+1}} \quad (5)$$

where $\frac{1}{C} = \int_{\mathcal{G}}\left[\left(\frac{\mathrm{d}\chi(t)}{\mathrm{d}t}\right)^{\kappa}\frac{\partial^{\kappa} f(\chi(t);t)}{\partial x^{\kappa}}\phi(t)\right]^{\frac{1}{\kappa+1}}\mathrm{d}t.$

Proof. See Appendix A. □

Although the optimal density is derived in the special case of scalar quantities and the HR regime, the corresponding result is insightful both for the analysis and the design. The conventional result when distortion minimization is pursued is that the optimal density $\rho^{\star}$ is proportional to $\phi^{\frac{1}{3}}(g)$. In practice this means allocating more quantization bits to more likely realizations of g. Under the GOQ approach, this conclusion is seen to be questioned. Indeed, the best density is seen to result from a combined effect of the parameter density ϕ, the variation speed of f w.r.t. the decision x (that is, the sensitivity of the goal regarding the decision), and the smoothness of the decision function χ w.r.t. the parameter to be quantized. As a consequence all these three factors need to be acccounted for in practice to design a good GOQ and allocate quantization bits in particular. Let us illustrate this with a simple example that is relevant to the problem of energy-efficient wireless transmit power control.

Example. Consider the following energy-efficiency (EE) performance metric $f(x;g) = -\frac{\exp\left(-\frac{c}{xg}\right)}{x^{\eta}}$ with $c > 0$ and $\eta \geqslant 2$. Here x represents the transmit power and g the channel gain [47] . Assume the channel gain g is exponentially distributed that is, $\phi(g) = \frac{1}{\overline{g}}\exp\left(-\frac{g}{\overline{g}}\right)$ with $\mathbb{E}(g) = \overline{g} > 0$. One obtains that $\kappa = 2$, $\chi(g) = \frac{c}{\eta g}$ and

$$\rho^{\star}(g) = C\left[\frac{\eta^{\eta+1}}{c^{\eta}e^{\eta}}g^{\eta-2}\phi(g)\right]^{\frac{1}{3}}. \quad (6)$$

For instance, for $\eta = 3$, it is easy to check that the quantization interval density $\rho^{\star}$ is increasing for $0 \leqslant g \leqslant \overline{g}$ then decreasing for $g \geqslant \overline{g}$. This result thus markedly differs from the conventional distortion-based approach. Indeed, under the latter approach, one would allocate more quantization bits to small values of the channel gain (since ϕ is strictly decreasing). Under the GOQ approach, most of the allocation bits should be allocated for values around the mean value of g.

In this section, we have been searching for the best scalar GOQ for a given goal function f. Now, we would like to provide some elements about the relationship between the nature of f and the quantization performance. For example, it is known that compressing a signal for which its energy is concentrated at small frequencies is generally an easy task. Similarly, here, we would like to know more about the connection between the regularity properties of the goal function and the level of difficulty to quantize its parameters. Since, this relevant issue constitutes a challenging mathematical problem, we only provide some preliminary results to explore this promising direction. For this purpose, we assume the chosen quantizer to be given by the optimal HR quantizer given by $\rho^{\star}$ and study the impact of f on $L(\rho^{\star};f)$. To be rigorous and clearly indicate the dependency of $\rho^{\star}$ regarding f, we will use the notation $\rho_f^{\star}$.

B. About choosing the scaling factor α_f

So far, since f was fixed, the scaling factor α_f in the definition of the OL L was not relevant. But when it comes to minimizing $L(\rho_f^{\star};f)$ w.r.t f, this factor plays an important role. Indeed, if one wants to compare the hardness to compress of two functions, the retained performance criterion has to possess some invariance properties. In particular, it should be invariant to affine transformations. The OL has not this property regarding f since a function of the form $F = Af+B$ (with $A > 0$) would produce a large OL when A is large even if the OL obtained for the original f is small. Hence the need for normalizing the OL properly and thus the presence of α_f. Here, we consider two choices for α_f, which amounts to considering two different reference case for the performance comparison. The first reference case is uniform quantization. For this case, the normalizing factor is denoted by α_f^{UQ} and chosen to be the reciprocal of the OL obtained when using a HR uniform quantizer (UQ). It expresses as:

$$\frac{1}{\alpha_f^{\mathrm{UQ}}} = \int_{\mathcal{G}} C_g^{-\kappa}\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa} f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g \quad (7)$$

where $\int_{\mathcal{G}} C_g \mathrm{d}g = 1$. This case allows one to quantify the potential gain from using a GOQ instead of a standard quantizer which is independent of the goal function. The second reference case we consider corresponds to the situation where the DM entity takes a constant decision (CD) independently of the value of g. This would correspond to the situation where no instantaneous information about g is available and only statistics can be exploited. Although this reference case is not necessarily the right benchmark for a given application it is still of interest for extracting useful insights because, this time, it is not about comparing two quantizers but more about measuring the intrinsic difficulty to compress a given function. By defining $\overline{x}$ the chosen constant decision as $\overline{x} \in \underset{x\in\mathcal{X}}{\arg\min}\ \mathbb{E}_g\left[f(x;g) - f(\chi(g);g)\right]$, the corresponding normalizing factor is denoted by α_f^{CD} and expresses as:

$$\frac{1}{\alpha_f^{\mathrm{CD}}} = \frac{1}{(2M)^{\kappa}\kappa!(\kappa+1)}\int_{g\in\mathcal{G}}\left[f(\overline{x};g) - f(\chi(g);g)\right]\phi(g)\,\mathrm{d}g \quad (8)$$

where $\overline{x}$ is the chosen constant decision. The above quantity represents the OL obtained when using the best CD multiplied par a term in κ which comes from the HR approximation (see App. A for more details).

C. On the impact of the goal function on the OL

Equipped with these two versions of the (normalized) OL, comparing different goal functions becomes a well posed problem. For this purpose, we have selected several functions

Goal function $f(x;g)$	**p.d.f.** $\phi(g)$	**ODF** $\chi(g)$	**OL** ($\alpha_f = \alpha_f^{\mathrm{UQ}}$)	**OL** ($\alpha_f = \alpha_f^{\mathrm{CD}}$)
$\frac{\log(1+10gx)}{x}$	uniform	$[1-\frac{1}{10g}]^+$	0.00399	0.0488
$\frac{\exp(-\frac{1}{gx})}{x}$	uniform	$\frac{1}{g}$	0.648	6.5943
$\frac{(1-\exp(-gx))^{10}}{x}$	uniform	$\frac{3.6150}{g}$	0.648	19.4565
$(x-g)^2$	uniform	g	1	24
$\frac{\log(1+10gx)}{x}$	exp	$[1-\frac{1}{10g}]^+$	0.0019	0.4859
$\frac{\exp(-\frac{1}{gx})}{x}$	exp	$\frac{1}{g}$	0.083	18.75
$\frac{(1-\exp(-gx))^{10}}{x}$	exp	$\frac{3.6150}{g}$	0.083	61.12
$(x-g)^2$	exp	g	0.24	48.50

TABLE I: Comparison of different goal functions

[47]–[49] that frequently appear in wireless resource allocation problems. For the selected functions, all quantities at hand can be expressed analytically and the integral associated with the OL can be computed. The obtained results appear in Table I. With the parameter space taken to be the interval $[0.1, 10]$, the table assumes two different choices for the p.d.f. ϕ, the uniform distribution and a truncated exponential distribution namely, $\phi(g) = \frac{\exp(-g)}{\int_{0.1}^{10} \exp(-x)\mathrm{d}x}$. The two columns providing the value of the OL allows one to establish some hierarchy between the selected functions. The obtained results suggest that logarithm-type goal functions provide a relatively small OL. These types of function would be qualified as easy to compress, which means for example that a rough description of the parameter is sufficient to take a good decision. Quantizing finely the parameter would lead to a waste of resources. This interpretation which is based on the HR analysis will be confirmed by simulations performed in arbitrary regimes. In a wireless system, this would e.g., mean that transmission rate-type performance metrics are not very sensitive to quantization noise and therefore a coarse feedback on CSI is suited to the goal. The table shows a different behavior for exponential-type functions, which are typically used to model energy-efficiency in wireless systems. These types of function require a more precise description of the function parameters (e.g., the CSI). Implementing the GOQ approach for such functions is seen to still provide a quite significant gain in terms of OL when compared to uniform quantization. We see that the HR analysis of the scalar quantization case provides useful insights that could be both used for an ad hoc design of a goal-oriented quantizer and deepened by considering more complex performance metrics.

IV. Vector GOQ: High resolution analysis and proposed quantization algorithm

A. High resolution analysis

As motivated in Sec. III, for some applications vector quantization is not used for reasons such as computational complexity. This is the case for instance for MIMO systems where the transfer channel matrix entries are quantized by a set of scalar quantizers. But, for optimality reasons or because of the definition of the quantization problem, vector quantization may be necessary. For instance, it is of high practical interest to be able to cluster series of the non-flexible electrical power consumption over one day for example [50] [51] [52], which leads to a sample dimension of $p = 48$ when the power signal is sampled every 30 minutes. By construction, this clustering problem is similar to a vector quantization problem for which one wants to create a certain number (M with our notation) of data subsets. For this specific problem one may want to fix M to a small number, say $M = 4$, and distinguish between 4 consumption behaviors.

For the scalar case, it has been seen that the HR regime allows to determine the best goal-oriented quantizer, which is fully characterized by the density function $\rho^\star$ (see (5)). However, in the vector case, even under the HR assumption, the problem remains challenging in general. This is one of the reasons why we resort to approximations. The full analytical characterization of the corresponding approximations is left as a relevant extension of the present work. The goal in this paper is threefold: to show how these approximations can be used for the quantizer design; to support the choices made by simulations performed with a low and moderate number of quantization bits; to focus on the potential gains that can be brought by the GOQ approach. One the main results of this section consists in providing an exploitable approximation of the OL in the vector case. This approximation will be directly exploited further in this section for the quantizer design part. The result is stated through the following proposition.

Proposition IV.1. *Assume $d \geqslant 1$, $p \geqslant 1$, and $\kappa = 2$. Assume f and χ twice differentiable. Denote by $\mathbf{H}_f(x;g)$ the Hessian matrix of f and denote by $\mathbf{J}_\chi(g)$ the Jacobian matrix of f evaluated for an optimal decision $\chi(g)$. In the regime of large M, the optimality loss function $L(Q;f)$ defined as in (2) can be approximated as follows:*

$$L(Q;f) = \underbrace{\alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} (g-z_m)^{\mathrm{T}} \mathbf{A}_{f,\chi}(g)(g-z_m)\phi(g)\mathrm{d}g}_{\hat{L}_M(Q;f)} + o(M^{-\frac{2}{p}}) \tag{9}$$

where $\mathbf{A}_{f,\chi}(g) = \mathbf{J}_\chi^{\mathrm{T}}(g)\mathbf{H}_f(\chi(g);g)\mathbf{J}_\chi(g)$. Additionally, by assuming the Gersho hypothesis [53] (see App. B), the above first order HR equivalent of L can be bounded as $L_M^{\min}((Q;f) \leqslant \widehat{L}_M(Q;f) \leqslant L_M^{\max}(Q;f)$ with

$$L_M^{\min}(Q;f) = \frac{p\mu_p}{2} M^{-\frac{2}{p}} \left(\int_{\mathcal{G}} (\lambda_{\min}(g;f)\phi(g))^{\frac{p}{p+2}} \mathrm{d}g \right)^{\frac{p+2}{p}} \tag{10}$$

$$L_M^{\max}(Q;f) = \frac{p\mu_p}{2} M^{-\frac{2}{p}} \left(\int_{\mathcal{G}} (\lambda_{\max}(g;f)\phi(g))^{\frac{p}{p+2}} \mathrm{d}g \right)^{\frac{p+2}{p}} \tag{11}$$

where: $\lambda_{\min}(g;f)$ (resp. $\lambda_{\max}(g;f)$) is the smallest (resp. largest) eigenvalue of $\mathbf{A}_{f,\chi}(g)$ and μ_p is the least normalized

moment of inertia of the p-dimensional tessellating polytope $\mathbb{T}_p$ defined by

$$\mu_p = \min_{\mathbb{T}_p, z} \frac{1}{p} \frac{1}{\mathrm{vol}(\mathbb{T}_p)^{1+2/p}} \int_{\mathbb{T}_p} \|g - z\|^2 \mathrm{d}g. \tag{12}$$

Proof. See Appendix B. □

The first-order equivalent in Prop. IV.1 is seen to depend on the matrix $\mathbf{A}_{f,\chi}(g)$. This matrix corresponds to the vector generalization of the product $\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^2 \frac{\partial^2 f(\chi(g);g)}{\partial x^2}$ that appears in the scalar case and shows how the OL is related to the regularity properties of the goal function f. For the conventional quantization approach ($f(x;g) = \|x-g\|^2$), one has merely that $\mathbf{A}_{f,\chi}(g) = \mathbf{I}$. Therefore in the HR regime, the structure of the equivalent shows that considering a general goal function f amounts to introducing an appropriate weighting matrix in the original distortion function. This matrix will be precisely used to derive an algorithm to compute a good vector GO quantizer that is tailored to the goal function.

The derived lower and upper bounds can be used both for characterizing the performance of a GOQ and for the quantizer design, which is explained at the end of this section. The bounds are tight in special cases such as when $p = 1$ (in which case $\mu_p = \frac{1}{12}$) and when $f(x;g) = \|x-g\|^2$ (with no restrictions on the dimensions d and p. Generally speaking, the gap between the two bounds is observed to be small when p is less or much less than d. Now if $p \geqslant d$, it can be seen that $\lambda_{\min}(g;f) = 0$ since the matrix $\mathbf{A}_{f,\chi}(g)$ is not full rank. As a consequence, the lower bound derived in (10) is not tight anymore. Hence, it is necessary to derive a tighter lower bound in this scenario. To this end, one can treat $\mathbf{J}_\chi(g)e_m$, with $e_m = \frac{g-z_m}{\|g-z_m\|}$, as a vector and thus $e_m^{\mathrm{T}}\mathbf{A}_{f,\chi}(g)\,e_m$ is minimized if and only if $\mathbf{J}_\chi(g)e_m$ is aligned with the eigenvector associated with the smallest eigenvalue of $\mathbf{H}_f(\chi(g);g)$. By denoting $\nu_{\min}(g;f)$ the smallest eigenvalue of $\mathbf{H}_f(\chi(g);g)$, the term $e_m^{\mathrm{T}}\mathbf{A}_{f,\chi}(g)\,e_m$ can be lower bounded by $\nu_{\min}(g;f)\mathsf{a}(\mathbf{J}_\chi(g))$, where $\mathsf{a}(\mathbf{J}_\chi(g))$ is the scalar factor between $\mathbf{J}_\chi(g)e_m$ and the smallest eigenvector of $\mathbf{H}_f(\chi(g);g)$. By replacing $\lambda_{\min}(g;f)$ with $\nu_{\min}(g;f)\mathsf{a}(\mathbf{J}_\chi(g))$, a new lower bound can be derived for the case where $p \geqslant d$. The proposed refinement procedure can also be used for the upper bound on the OL but note that the upper bound is mainly dependent on p and is much less dependent on the dimensionality d, which makes the corresponding refinement generally less useful.

B. Proposed quantization algorithm

As mentioned in the last subsection, the bounds provided by Prop. IV.1 can be used to characterize the performance of a quantizer and study, at least numerically, the impact of the nature of f on the OL. In the present subsection, the main objective is to exploit the HR equivalent of Prop. IV.1 to design a practical quantization algorithm. Considering the fact that the optimal decision function may produce solution at the boundary of the decision set and that only sub-optimal decision function may be available in real systems, we relax here the optimality first order condition $\frac{\partial f(x;g)}{\partial x}|_{x=\chi(g)} = 0$. Therefore, the optimality loss can be written for algorithmic purposes in a more general form:

$$\begin{aligned} &L(Q;f) \\ &= \sum_{m=1}^{M} \int_{\mathcal{G}_m} \left[\left(\frac{\partial f(x;g)}{\partial x}\,|_{x=\chi(g)} \right)^{\mathrm{T}} (\chi(z_m) - \chi(g)) \right. \\ &\quad \left. + \frac{1}{2} (\chi(z_m) - \chi(g))^{\mathrm{T}} \mathbf{H}_{f,\chi}(g) (\chi(z_m) - \chi(g)) \right] \phi(g)\mathrm{d}g \\ &\quad + o\left(\|\chi(z_m) - \chi(g)\|^2\right) \end{aligned} \tag{13}$$

where $(\mathbf{H}_{f,\chi}(g))_{i,j} = \frac{\partial^2 f(x;g)}{\partial x_i \partial x_j}|_{x=\chi(g)}$ for $1 \leqslant i,j \leqslant p$. By using the Taylor expansion, we have that:

$$\begin{aligned} &\chi(z_m) - \chi(g) \\ =&\mathbf{J}_\chi(g)(z_m - g) + \begin{bmatrix} (z_m-g)^{\mathrm{T}}\mathbf{H}_{\chi_1}(g)(z_m-g) \\ (z_m-g)^{\mathrm{T}}\mathbf{H}_{\chi_2}(g)(z_m-g) \\ \cdots \\ (z_m-g)^{\mathrm{T}}\mathbf{H}_{\chi_d}(g)(z_m-g) \end{bmatrix} \\ &+o\left(\|z_m - g\|^2\right) \end{aligned} \tag{14}$$

where $(\mathbf{H}_{\chi_i}(g))_{\ell,k} = \frac{\partial^2 \chi_i(g)}{\partial g_\ell \partial g_k}$ for $1 \leqslant l,k \leqslant p$ and $1 \leqslant i \leqslant d$, $\chi(g) = [\chi_1(g), \ldots, \chi_d(g)]^{\mathrm{T}}$. Plugging this expression in the expression of $L(Q;f)$, the optimality loss can be re-expressed as

$$\begin{aligned} &L(Q;f) \\ =&\frac{1}{2} \sum_{m=1}^{M} \int_{\mathcal{G}_m} (g - z_m)^{\mathrm{T}} \mathbf{B}_{f,\chi}(g)(g - z_m)\phi(g)\mathrm{d}g \\ +&\frac{1}{2} \sum_{m=1}^{M} \int_{\mathcal{G}_m} (g - z_m)^{\mathrm{T}} \mathbf{A}_{f,\chi}(g)(g - z_m)\phi(g)\mathrm{d}g + o\left(M^{\frac{-2}{p}}\right) \end{aligned} \tag{15}$$

where $\mathbf{B}_{f,\chi}(g) = \sum_{i=1}^{d} \bigtriangledown f_i(g) \mathbf{H}_{\chi_i}(g)$ with

$$\frac{\partial f(x;g)}{\partial x}|_{x=\chi(g)} = (\bigtriangledown f_1(g), \bigtriangledown f_2(g), \ldots, \bigtriangledown f_d(g))$$

and $\mathbf{A}_{f,\chi}(g) = \mathbf{J}_\chi^{\mathrm{T}}(g)\mathbf{H}_f(x;g)\mathbf{J}_\chi(g)$.

By using this new expression of the OL, one exhibits a natural structure for applying an alternating optimization algorithm and thus to minimize $\widetilde{L} = \sum_{m=1}^{M} \int_{\mathcal{G}_m} (g - z_m)^{\mathrm{T}} (\mathbf{B}_{f,\chi}(g) + \mathbf{A}_{f,\chi}(g)) (g - z_m)\phi(g)\mathrm{d}g$ as follows:

- *Representative updating step*: To minimize $\tilde{L}$ with **fixed regions**, the problem boils down to find the representative z_m such that $\int_{\mathcal{G}_m} (g - z_m)^{\mathrm{T}} (\mathbf{B}_{f,\chi}(g) + \mathbf{A}_{f,\chi}(g)) (g - z_m)\phi(g)\mathrm{d}g$ can be minimized. One can apply a gradient descent technique to achieve that where the gradient can be easily found:

$$\frac{\partial \tilde{L}}{\partial z_m} = 2 \int_{\mathcal{G}_m} \mathbf{E}_{f,\chi}(g) (g - z_m) \phi(g) \mathrm{d}g \tag{16}$$

where $\mathbf{E}_{f,\chi}(g) = \mathbf{B}_{f,\chi}(g) + \mathbf{A}_{f,\chi}(g)$.

Algorithm 1: Goal-oriented Quantization Algorithm

Inputs: goal function $f(x;g)$, $\chi(g)$, error tolerance ε, number of cells M and number of iterations T;
Inputs: $\mathcal{Z}^{(0)} = \left\{z_1^{(0)}, \ldots, z_M^{(0)}\right\}$;
Inputs: $\mathcal{G}^{(0)} = \left\{\mathcal{G}_1^{(0)}, \ldots, \mathcal{G}_M^{(0)}\right\}$;
for $t = 1$ **to** T **do**
  for $m = 1$ **to** M **do**
    Update $\mathcal{G}_m^{(t)}$ by $\left\{g \middle| \widetilde{\ell}_f\left(g, z_m^{(t-1)}\right) \leqslant \widetilde{\ell}_f\left(g, z_{m'}^{(t-1)}\right), \forall m' \neq m\right\}$;
    Update $z_m^{(t)}$ by $z_m^{(t)} = z_m^{(t-1)} - r_t \frac{\partial \widetilde{L}\left(\mathcal{Z}^{(t-1)}\right)}{\partial z_m^{(t-1)}}$ with the step size $r_t > 0$ s.t. $z_m^{(t)} \in \mathcal{G}$;
  end
  if $\sum_{m=1}^{M} \left\| z_m^{(t)} - z_m^{(t-1)} \right\|^2 < \varepsilon$ **then**
    Break;
  end
end
Outputs: $\mathcal{Z}^\star = \mathcal{Z}^{(t)}$ and $\mathcal{G}^\star = \mathcal{G}^{(t)}$;

- *Region updating step*: For given representatives, the region can be computed as:

$$\begin{aligned}\mathcal{G}_m = \ & \left\{g | (g - z_m)^{\mathrm{T}} \mathbf{E}_{f,\chi}(g)(g - z_m)\right. \\ & \left. \leqslant (g - z_{m'})^{\mathrm{T}} \mathbf{E}_{f,\chi}(g)(g - z_{m'})\right\}\end{aligned}$$

where $m' \neq m$.

The approximate individual optimality loss is thus defined by $\widetilde{\ell}_f(g, z)$ of the parameter g w.r.t. a representative z as:

$$\widetilde{\ell}_f(g, z) \triangleq (g - z)^{\mathrm{T}} \mathbf{E}_{f,\chi}(g)(g - z). \tag{17}$$

our goal-oriented quantization algorithm is summarized in pseudo-code form through algorithm 1. The proposed algorithm can be applied to the scalar case. In the latter case, the matrix $\mathbf{A}_{f,\chi}(g)$ becomes $\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^2 \frac{\partial^2 f(\chi(g);g)}{\partial x^2}$ which corresponds to the term appearing in Equation 5 with $\kappa = 2$. And we have that $\mathbf{B}_{f,\chi}(g) = 0$. The reason for this is that either the first-order optimality condition holds or the lower and upper bounds of the quantization interval are fixed points.

V. Numerical performance analysis

In this section we both want to illustrate some analytical results derived in the preceding sections and also see, from purely numerical results, to what extent some insights obtained from the HR analysis hold in scenarios where main assumptions such as smoothness are relaxed. For this purpose, we consider four goal functions: an exponential-type goal function and a log-type goal function which are relevant for GO information quantization problems in wireless resource allocation problems; a quadratic-type goal function which is typically relevant for GOQ in controlled systems; an $\mathrm{L_P}$ norm-type goal function which is relevant for GO data clustering/quantization in power systems.

A. Impact of the goal function on the OL for wireless metrics

Table I provides analytical results for the scalar case in the HR regime. It suggests that for a given quantization scheme, log-type goal functions lead smaller values for the OL than exp-type goal functions. Let us consider the performance metric introduced by [48] to measure the EE of a multiband communication: $f^{\mathrm{EE}}(x;g) = -\frac{\sum_{i=1}^{S} \exp\left(-\frac{c}{x_i g_i}\right)}{\sum_{i=1}^{S} x_i}$ where S is the number of bands, $c > 0$, x_i is the transmit power for band s, and g_i the channel gain for band s. The log-type function is taken to be the classical spectral efficiency (SE) function $f^{\mathrm{SE}}(x;g) = -\sum_{i=1}^{S} \log\left(1 + x_i \frac{g_i}{\sigma^2}\right)$. We impose that $x_i \geqslant 0$ and $\sum_{i=1}^{S} x_i \leqslant P_{\max}$. For $\frac{P_{\max}}{\sigma^2} = 5$, $c = 1$, $S = 2$, and a uniform quantizer Fig. 2 depicts the relative OL in percentage (relatively to the ideal case):

$$\text{Relative OL}(\%) = 100 \times \left(\frac{f(\chi(Q(g)); g) - f(\chi(g); g)}{f(\chi(g); g)}\right) \tag{18}$$

averaged over 10000 independent Rayleigh fading realizations (with $\mathbb{E}(g) = 1$) against the number of quantization bits per realization of g. We see for a given number of bits per sample, the OL for the SE function is much smaller than the SE function. We retrieve the hierarchy suggested by Table I. This shows that the SE function can accommodate a rough quantization of the parameters (that is, the channel gains) without degrading significantly the DM process, which is to choose a good power allocation vector. Using a fine quantizer would lead to waste of resources for the SE function (here we see that a 1-bit quantizer yields an OL of about 2%, which illustrates well the importance of adapting the quantizer to the goal function.

B. Performance gains obtained from tailoring the quantizer to the (control) goal

Now we assume $d = p = 2$ and consider the following quadratic function:

$$f^{\mathrm{QUA}}(x;g) = (x_1 - h_1(g))^2 + (x_2 - h_2(g))^2 + (x_1 - x_2)^2 \tag{19}$$

with $h_1(g) = 2g_1g_2 - \frac{1}{2}g_1^2g_2^2$ and $h_2(g) = g_1^2g_2^2 - g_1g_2$. Parameters are assumed to be i.i.d. and exponentially distributed, i.e., $\phi(g) = \exp(-g_1 - g_2)$. One can check that $\chi(g) = [g_1g_2, \frac{1}{2}g_1^2g_2^2]^{\mathrm{T}}$. In Fig. 3, the relative OL in percentage (relatively to the ideal case) against the number of regions M is represented for a conventional vector quantizer (namely, a distortion-based quantizer implementing the Lloyd-Max algorithm [54], [55]), hardware-limited task-based quantization (HLTB) in [17] and for the proposed vector GOQ computed thanks to algorithm 1. Although Algorithm 1 is based on a HR approximation, it is seen to provide a very significant gain in terms of OL even for a small number of regions. For $M = 5$ a conventional quantizer would lead to a relative OL of 70% which is a significant performance degradation w.r.t. the ideal case where g is perfectly known,

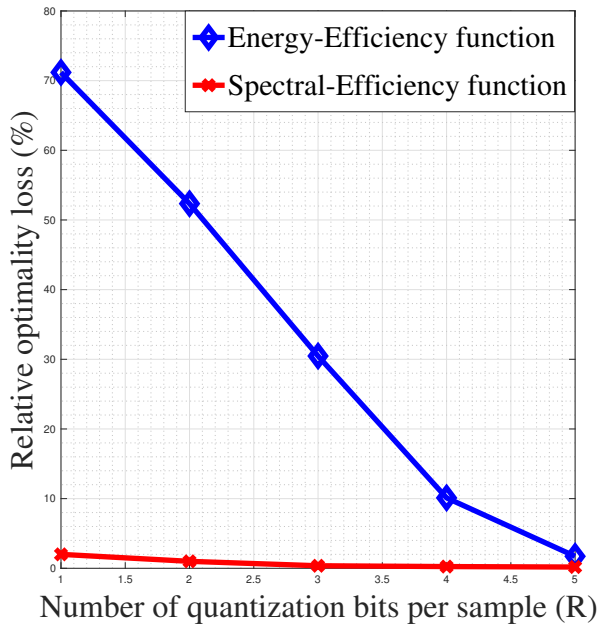


Fig. 2: The figure shows the impact of the number of quantization bits on the decision-making quality (measured in terms of optimality loss) on two different well-used goal functions. Log-type SE functions appear to accommodate very well with very rough quantization for its parameters (CSI) which is the not the case for exp-type EE functions. This simulation is in accordance with the analytical results of Table I.

whereas the proposed GOQ allows the OL to be as low as 10%. Besides, compared to HLTB quantizer which is also goal-oriented, the optimality loss reduction of proposed algorithm is still considerable in low-resolution regime. The explanation behind this performance gain is already available through Example 1 in which we have seen the importance of adapting the "density" or more generally the concentration of the regions (and thus allocating the quantization bits) not according to the parameter distribution (conventional approach) but to an appropriately weighted distribution. This difference is illustrated through Fig. 5. The top subfigure shows the p.d.f. of the parameter g (namely $\phi(g)$). The bottom subfigure shows $\lambda_{\max}\left(g; f^{\mathrm{QUA}}\right)\phi\left(g\right)$. The analysis conducted in Sec. III suggests to concentrate the quantization regions according to this weighted density, which is markedly different from ϕ. By doing so, Algorithm 1 provides a very significant improvement, the main powerful insight being not to allocate quantization resources to the most likely realizations of the information source but to the ones that impact the most the goal, which is measured through the weighted density $\lambda_{\max}\left(g; f^{\mathrm{QUA}}\right)\phi\left(g\right)$. Notice that the above numerical results are obtained when the p.d.f. of g is known. In practice, it might happen that this p.d.f. is not available or is time-varying. Then one can easily adapt algorithm 1 by replacing statistical means with empirical/sample means and possibly, refreshing

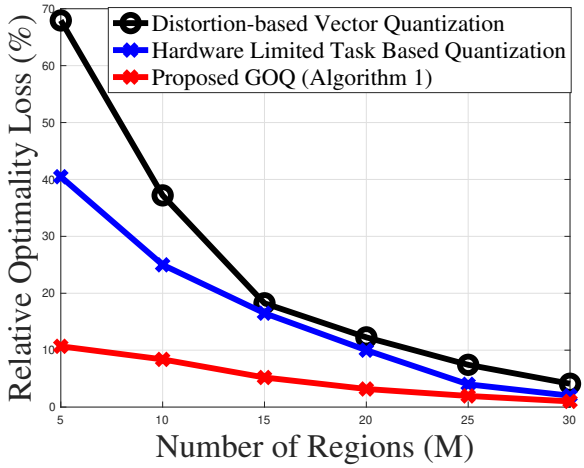


Fig. 3: The goal function being a quadratic function, the figure shows the importance in terms of (decision) optimality loss of adapting the quantizer to the goal instead of using the conventional distortion-based quantization approach.

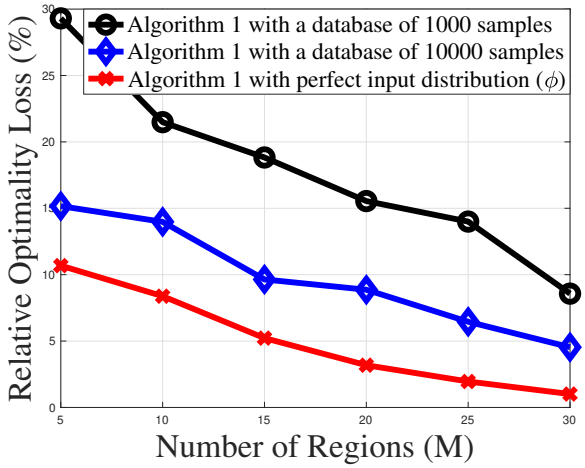
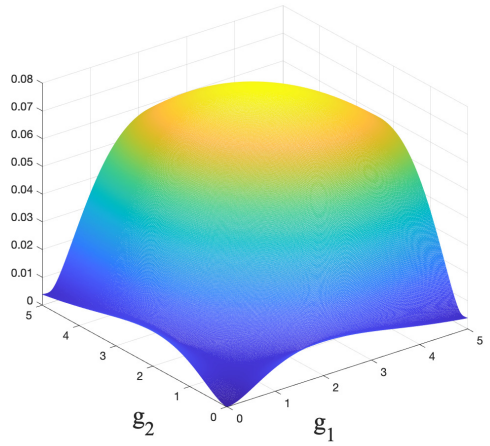


Fig. 4: The figure assesses the performance loss due to not knowing the input distribution ϕ perfectly but rather with a low number of samples took from a database.

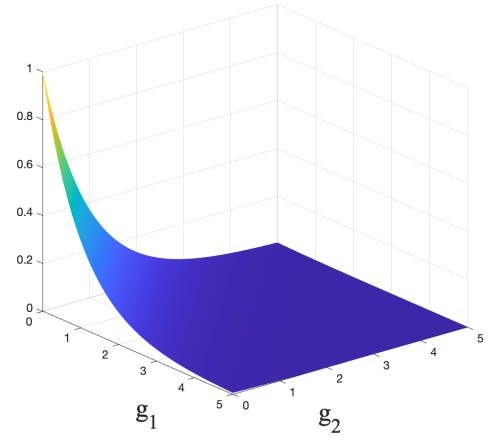
the database on the fly if the statistics need to be tracked. Fig. 4 precisely shows the loss that would be induced by using a relatively small database instead of knowing the input distribution perfectly. One can observe that the data-based GO quantizer still could achieve a relative optimality loss of 9% for a database with only 1000 data points, which illustrates the relevance of the proposed method when the input distribution is not available.

C. Goal-oriented quantization and power consumption scheduling

‣ Now we assume $d = p = 24$. We consider a performance metric which is relevant for a communication problem in



(a) New density $\lambda_{\max}\left(g; f^{\mathrm{QUA}}\right)\phi\left(g\right)$ of GOQ algorithm



(b) Original probability distribution $\phi\left(g\right)$

Fig. 5: The figure shows the marked difference between the parameter probability distribution (bottom curve) and the probability distribution of interest that is relevant to the decision-making task (top curve). It implies in particular that quantization regions (and thus quantization bits) should be allocated in a very different way from the conventional way.

the smart grid. Indeed, we consider that the goal function $f^{\mathrm{PCS}}(x;g) = \|x+g\|_{\mathrm{P}}$, P being the exponent power parameter of the $\mathrm{L_P}$ norm, and PCS stands for power consumption scheduling. This time the vector $x = (x_1, ..., x_d)$ ($d = p$ here) represents the chosen flexible power consumption scheduling strategy; we impose that $x_i \geqslant 0$ and $\sum_{i=1}^{d} x_i \geqslant E$, $E > 0$ being the desired energy level chosen as 30 kWh in our simulation setting. The parameter vector g represents the non-controllable part of the power. When P becomes large, the problem amount to limiting the peak power. The clustering problem is a data-based counterpart of the quantization problem in which a finite set of realizations for g is available (instead of the knowledge of ϕ). We want to cluster a finite dataset into clusters or groups of data (instead of

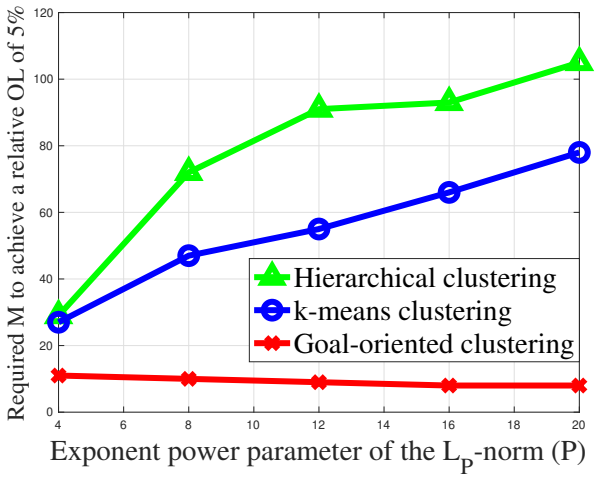


Fig. 6: Required number of clusters (M) against the Exponent power parameter of the $\mathrm{L_P}$-norm (P) for the k-means and goal-oriented clustering. The goal-oriented clustering approach yields a drastic reduction in terms of the number of clusters when P increases.

continuous regions). And the goal is to minimize f^{PCS} by only having a clustered version of the data. For the purpose of applying the GOQ approach to clustering, we make the following two choices in terms of implementation. First, the statistical expectation is replaced with its empirical version in the algorithm; the empirical mean is performed over the 300 time series of the Pecanstreet dataset. Second, since the number of samples is small, representatives are computed by directly minimizing $L(Q, f)$ (as in [52]) instead of the approximated version $\widetilde{L}$. For a given relative OL of 5% one then looks at the number of required clusters (that is, M) versus the exponent power parameter of the $\mathrm{L_P}$ (that is, P). In Fig. 6, we compare the performance of the the GO clustering technique with the $k-$means algorithm (which is exactly the data-based counterpart of the LM algorithm) and hierarchical clustering (HC) algorithm for the Pecanstreet database [56]. For HC, the squared Euclidean distance and weighted pair group method with arithmetic mean are used. First one can observe that partitioning clustering slightly outperforms hierarchical clustering, this might be explained by the fact that several clusters in HC compose of a single outlier data point (in terms of Euclidean distance), but outlier data points might yield similar decision as normal data points for Lp-norm problems especially with large p. For P ranging from 4 to 20, the figure shows that the number of required clusters can be decreased from about $M = 80$ to $M = 8$ by adapting the clustering technique to the final decision instead of creating clusters based on an exogenous similarity index, which is the Euclidean norm in the case of the $k-$means algorithm.

VI. Conclusion

In this paper, the focus is on one key element of a goal-oriented communication chain namely, the quantization stage.

The GOQ problem is very relevant for lossy data compression e.g., to have high spectral efficiency in wireless systems (by transmitting only the minimum amount of information relevant to the correct task execution). It is also relevant for many resource allocation problems, hence the choices for the goal function in this paper. One of the contributions of this paper is to exploit the HR assumption both for the analysis and design of a GOQ. Valuable insights of practical interest have been obtained. Let us mention two of them. The most conventional way of designing a source coder is to allocate resources (say bits) according to the frequency of the realization of the source symbol (this is what Huffman and arithmetic coding schemes and their many variants do). Our analysis shows that this approach may lead to a significant performance degradation and rather shows in a precise way (see e.g., Prop. III.1, Example 1, and Fig. 4) how the variation speed of the goal and decision functions should be taken into account to allocate such resources in a much more efficient way. Our analysis also allows one to make progresses into the direction of understanding how the goal function impacts the quantizer. Both analytical and simulation results are provided to exhibit the existence of possible classes of functions which would more or less easy to be compressed. This knowledge allows the quantizer to be matched to the goal. For example, rough quantization seem to have a small impact on the task execution as far as log-type goal functions are concerned. The behavior is different for exp-type functions. This suggests for example that CSI feedback should be much finer for energy-efficient performance metrics than for spectral-efficiency metrics. It is seen that the proposed framework is rich in terms of practical insights. Nonetheless, many relevant issues are left open and would need to be explored. For instance, theoretical analysis relies on smoothness assumptions for the goal and decision functions. What would the results become for non-smooth functions? The functions are also assumed to be known. How to adapt the approach when only the realizations of these functions are available? Also a dedicated complexity analysis should be conducted. Generally, the problem of designing vector GO quantizers when the dimension increases is open. An interesting extension of this work would also be to address the case of a non-stationary source, leading to the problem of an adaptive quantizer. How learning techniques could be used to solve all these issues?

APPENDIX A
PROOF OF PROPOSITION III.1

By using Taylor expansion, the optimality loss in high-resolution regime can be approximated by

$$\begin{aligned}
&L\left(Q;f\right)\\
=&\alpha_f \sum_{m=1}^{M}\int_{\mathcal{G}_m}\left[f\left(\chi\left(z_m\right);g\right)-f\left(\chi\left(g\right);g\right)\right]\phi\left(g\right)\mathrm{d}g\\
\overset{(a)}{=}&\alpha_f \sum_{m=1}^{M}\int_{\mathcal{G}_m}\left(\chi\left(z_m\right)-\chi\left(g\right)\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(x;g)}{\partial x^{\kappa}}|_{x=\chi(g)}\phi(g)\mathrm{d}g+o\left(M^{-\kappa}\right)\\
\overset{(b)}{=}&\alpha_f \sum_{m=1}^{M}\int_{\mathcal{G}_m}\left(z_m-g\right)^{\kappa}\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g+o\left(M^{-\kappa}\right)\\
\overset{(c)}{=}&\alpha_f \int_{\mathcal{G}}\frac{\Delta^{\kappa}(g)}{(\kappa+1)2^{\kappa}}\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g+o\left(M^{-\kappa}\right)\\
\overset{(d)}{=}&\frac{\alpha_f}{(2M)^{\kappa}(\kappa+1)!}\int_{\mathcal{G}}\rho^{-\kappa}(g)\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g+o\left(M^{-\kappa}\right)
\end{aligned} \tag{20}$$

(a) corresponds to the Taylor expansion of $(f(\chi(z_m);g)-f(\chi(g);g))$ in the regime of large M (infinitesimals of $M^{-\kappa}$ are not considered further); (b) follows from the fact that the higher order terms in the Taylor expansion of $(\chi(z_m)-\chi(g))$ are negligible w.r.t. the first term. (c) extends the idea of approximating mean-square error distortion in high resolution regime (see [57], [58]) to cases with even-order κ, i.e.,

$$\begin{aligned}
&\int_{\mathcal{G}_m}\left(z_m-g\right)^{\kappa}\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g\\
\approx&\left(\frac{\mathrm{d}\chi(z_m)}{\mathrm{d}z_m}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(z_m);z_m)}{\partial x^{\kappa}}\phi(z_m)\int_{z_m-\frac{\Delta(z_m)}{2}}^{z_m+\frac{\Delta(z_m)}{2}}\left(z_m-g\right)^{\kappa}\mathrm{d}g\\
\approx&\left(\frac{\mathrm{d}\chi(z_m)}{\mathrm{d}z_m}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(z_m);z_m)}{\partial x^{\kappa}}\phi(z_m)\frac{\Delta(z_m)^{\kappa}}{(\kappa+1)2^{\kappa}}\Delta(z_m)\\
\approx&\int_{\mathcal{G}_m}\frac{\Delta^{\kappa}(g)}{(\kappa+1)2^{\kappa}}\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{1}{\kappa!}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\mathrm{d}g;
\end{aligned} \tag{21}$$

(d) follows from results on high resolution quantization referring to equation (3). After the derivation optimality loss with high-resolution quantization theory, we aim to find the optimal quantization point density to minimize the OL. We first introduce a new function called value density:

$$p(g)=\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa}f(x;g)}{\partial x^{\kappa}}|_{x=\chi(g)}\phi(g)\geqslant 0. \tag{22}$$

Then we resort to the Hölder's inequality:

$$\int p^{\frac{1}{\kappa+1}}\leqslant\left(\int p\rho^{-\kappa}\right)^{\frac{1}{\kappa+1}}\left(\int\rho\right)^{\frac{\kappa}{\kappa+1}} \tag{23}$$

knowing $\left(\int\rho\right)^{\frac{\kappa}{\kappa+1}}=1$, it can be inferred that $\int p\rho^{-\kappa}\geqslant\left(\int p^{\frac{1}{\kappa+1}}\right)^{\kappa+1}$, with equality if and only if $p\rho^{-\kappa}=C_1\rho$ with $C_1>0$. The optimum density function of quantization points can thus be written as:

$$\rho^{\star}(g)=\frac{\left[\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\right]^{\frac{1}{\kappa+1}}}{\int_{\mathcal{G}}\left[\left(\frac{\mathrm{d}\chi(t)}{\mathrm{d}t}\right)^{\kappa}\frac{\partial^{\kappa}f(\chi(t);t)}{\partial x^{\kappa}}\phi(t)\right]^{\frac{1}{\kappa+1}}\mathrm{d}t}. \tag{24}$$

By plugging the optimal density into the expression of the optimality loss, when M is large, the OL $L\left(Q;f\right)$ becomes:

$$\begin{aligned}
&\lim_{M\to\infty}L\left(Q;f\right)\\
=&\frac{\alpha_f}{(2M)^{\kappa}(\kappa+1)!}\left(\int_{\mathcal{G}}\left[\left(\frac{\mathrm{d}\chi(g)}{\mathrm{d}g}\right)^{\kappa}\frac{\partial^{\kappa}f(\chi(g);g)}{\partial x^{\kappa}}\phi(g)\right]^{\frac{1}{\kappa+1}}\mathrm{d}g\right)^{\kappa+1}
\end{aligned} \tag{25}$$

APPENDIX B
PROOF OF PROPOSITION IV.1

To facilitate the derivation, we introduce the multi-index notation in order to represent partial derivative of the goal function. The d-dimensional multi-index can be written as $n = (n_1, \ldots, n_d)$. Its sum and factorial can be expressed as $|n| = \sum_{t=1}^{d} n_t$ and $n! = \prod_{t=1}^{d} n_t!$, respectively. Considering the decision variable $x = (x_1, \ldots, x_d)$, the partial derivative with degree n w.r.t. x can be expressed as $\mathfrak{D}_x^n f = \frac{\partial^{|n|} f}{\partial x_1^{n_1} \ldots \partial x_d^{n_d}}$, and the multi-index power of x can be written as $x^n = \prod_{i=1}^{d} x_i^{n_i}$.

By using the Taylor expansion for multivariate functions, the optimality loss can be rewritten as:

$$\begin{aligned} & L(Q;f) \\ =& \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} \left[f(\chi(z_m);g) - f(\chi(g);g) \right] \phi(g) \mathrm{d}g \\ =& \sum_{m=1}^{M} \left[\sum_{n:|n|\leqslant\kappa} \int_{\mathcal{G}_m} \frac{\mathfrak{D}_x^n f(\chi(g);g)}{n!} \left(\chi(z_m) - \chi(g)\right)^n \phi(g) \mathrm{d}g \right. \\ & \left. + \sum_{\hat{n}:|\hat{n}|=\kappa+1} \int_{\mathcal{G}_m} O\left(\left(\chi(z_m) - \chi(g)\right)^{\hat{n}} \right) \phi(g) \mathrm{d}g \right] \end{aligned} \tag{26}$$

Interestingly, one can note that the $\frac{\mathfrak{D}_x^n f(\chi(g);g)}{n!}$ are the components of the gradient vector of f w.r.t. x when $|n| = 1$, and $\frac{\mathfrak{D}_x^n f(\chi(g);g)}{n!}$ are the components of the Hessian matrix of f w.r.t. x when $|n| = 2$. For the terms with $|n| \geqslant 3$, it could be seen as the infinitesimal of the second order terms. Therefore, we could take $k = 2$ and ignore the higher order terms in high resolution regime. In addition, here we consider the scenario where the optimal decision function $\chi(.)$ always locates in the interior of the feasible set $\mathcal{X}$, and thus each component of the gradient vector is zero, namely, $\frac{\partial f(x;g)}{\partial x_t}|_{x=\chi(g)} = 0$. The optimality loss can be approximated by:

$$\begin{aligned} & L(Q;f) \\ =& \underbrace{\sum_{m=1}^{M} \sum_{n:|n|=2} \int_{\mathcal{G}_m} \frac{\mathfrak{D}_x^n f(\chi(g)\, g)}{n!} \left(\chi(z_m) - \chi(g)\right)^n \phi(g) \mathrm{d}g}_{\widehat{L}_M(Q;f)} + o\left(M^{-\frac{2}{p}}\right) \end{aligned} \tag{27}$$

and the $\widehat{L}_M(Q;f)$ can be further simplified as

$$\begin{aligned} & \widehat{L}_M(Q;f) = \\ \overset{(a)}{=}& \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} \frac{1}{2} (\chi(z_m) - \chi(g))^{\mathrm{T}} \mathbf{H}_f(\chi(g);g) (\chi(z_m) - \chi(g)) \phi(g) \mathrm{d}g \\ \overset{(b)}{=}& \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} \frac{1}{2} (\mathbf{J}_\chi(g)(z_m - g))^{\mathrm{T}} \mathbf{H}_f(\chi(g);g) (\mathbf{J}_\chi(g)(z_m - g)) \phi(g) \mathrm{d}g \\ \overset{(c)}{=}& \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} \frac{1}{2} \|g - z_m\|_2^2 e_m^{\mathrm{T}} \mathbf{J}_\chi^{\mathrm{T}}(g) \mathbf{H}_f(\chi(g);g) \mathbf{J}_\chi(g)\, e_m \phi(g) \mathrm{d}g \end{aligned} \tag{28}$$

where e_m is defined as the normalized vector of the difference, i.e., $e_m = \frac{g - z_m}{\|g - z_m\|_2}$. (a) follows from the fact that the second order term in the Taylor expansion can be rewritten with matrix multiplication using Hessian matrix; (b) follows from the fact that the higher order term in the Taylor expansion of $(\chi(g) - \chi(g_m))$ are negligible w.r.t. the first order term; (c) can be verified by defining e_m. It is worth noting that this expression is similar to the classical vector quantization while the p.d.f. of g is weighted by a new coefficient related to the Hessian and Jacobian of the goal function and the normalized vector e_m. To simplify the formula, we denote by $\mathbf{A}_{f,\chi}(g) = \mathbf{J}_\chi^{\mathrm{T}}(g) \mathbf{H}_f(\chi(g);g) \mathbf{J}_\chi(g)$, then one has that:

$$\widehat{L}_M(Q;f) = \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} \frac{1}{2} \|g - z_m\|_2^2 e_m^{\mathrm{T}} \mathbf{A}_{f,\chi}(g)\, e_m \phi(g) \mathrm{d}g. \tag{29}$$

As the normalized vector e_m depends both on g and the representative z_m, the vector case can not be tackled as the scalar case. Nevertheless, we will show similar properties could be found in the vector case. To directly approximate the OL defined in (28) is complicated, we thus resort to some matrix properties to bound OL. The accuracy of our approximation depends on how we approximate the term $e_m^{\mathrm{T}} \mathbf{A}_{f,\chi}(g)\, e_m$. For a given parameter g, maximum eigenvalue and minimum eigenvalue of matrix $\mathbf{A}_{f,\chi}(g)$ are denoted by $\lambda_{\max}(g;f)$ and $\lambda_{\min}(g;f) \geqslant 0$ respectively since the Hessian matrix $\mathbf{H}_f(\chi(g);g)$ is nonnegative definite due to optimum. Therefore, the term $e_m^{\mathrm{T}} \mathbf{A}_{f,\chi}(g)\, e_m$ can be upper bounded by $\lambda_{\max}(g;f)$ and lower bounded by $\lambda_{\min}(g;f)$.

We first study the lower bound of $\widehat{L}_M(Q;f)$. Similarly, we extend the notation of the point density $\rho(g)$ to a vector case which determines the approximate fraction of representatives contained in that region. Define the normalized moment of inertia of the cell $\mathcal{G}_m$ with representative z_m by

$$\mathscr{M}(\mathcal{G}_m, z_m) = \frac{1}{p} \frac{1}{\mathrm{vol}(\mathcal{G}_m)^{1+2/p}} \int_{\mathcal{G}_m} \|g - z_m\|_2^2 \mathrm{d}g, \tag{30}$$

and the inertial profile $\mathfrak{m}(g) = \mathscr{M}(\mathcal{G}_m, z_m)$ when $g \in \mathcal{G}_m$, the OL can be further approximated as [53] [12]:

$$\begin{aligned} & \mathrm{L}(Q;f) \\ =& \alpha_f \sum_{m=1}^{M} \int_{\mathcal{G}_m} (f(\chi(z_m);g) - f(\chi(g);g)) \phi(g) \mathrm{d}g \\ \overset{(a)}{\geqslant}& \alpha_f \sum_{m=1}^{M} \int_{g \in \mathcal{G}_m} \frac{1}{2} \|g - z_m\|_2^2 \lambda_{\min}(g;f) \phi(g) \mathrm{d}g \\ \overset{(b)}{=}& \sum_{m=1}^{M} \frac{\alpha_f p}{2M^{2/p}} \frac{\mathscr{M}(\mathcal{G}_m, z_m)}{\rho^{2/p}(z_m)} \lambda_{\min}(z_m;f) \phi(z_m) \mathrm{vol}(\mathcal{G}_m) \\ \overset{(c)}{=}& \frac{\alpha_f p}{2M^{2/p}} \int \frac{\mathfrak{m}(g)}{\rho^{2/p}(g)} \lambda_{\min}(g;f) \left(\phi(g)\right) \mathrm{d}g \end{aligned} \tag{31}$$

(a) comes from the fact that e_m is a normalized vector; (b) uses the definition of $\mathscr{M}(\mathcal{G}_m, z_m)$ and the relation $\lim_{M\to\infty} \sum_{m=1}^{M} \mathrm{vol}(\mathcal{G}_m) \rho(z_m) = M$; (c) is still the definition of Riemman integral. This result can be seen as a special case of Bennett's integral (see [57] [12]) by replacing $\phi(g)$ by the product $\lambda_{\min}(g;f)\phi(g)$. However, it is not known how to find the optimal inertial profile $\mathfrak{m}(g)$ and it is not even known what functions are allowable as inertial profiles. To this end, Gersho

[53] made the widely accepted hypothesis or conjecture that when R is large, most regions of a p-dimensional quantizer aims at minimizing or nearly minimizing the mean square error are approximately congruent to some basic tessellating p-dimensional cell shape $\mathbb{T}_p$. With this conjecture, the optimal inertial profile $\mathfrak{m}(g)$ can be seen as a constant μ_p in high resolution case. By using the Hölder's inequality, the optimal density $\rho(g)$ that minimizes the distortion can be written as

$$\rho^{\star}(g) = \frac{(\lambda_{\min}(g;f)\phi(g))^{\frac{p}{p+2}}}{\int_{\mathcal{G}} (\lambda_{\min}(t;f)\phi(t))^{\frac{p}{p+2}}\,\mathrm{d}t} \tag{32}$$

resulting in the low bound of distortion in (10). The same reasoning can be applied to the derivation of the proposed upper bound.

Remark When the number of cells is large, one has that $\mathfrak{m}(z_m) \approx \mathfrak{m}(g)$. Then one is able to define the inertial profile $\mathfrak{m}(g)$ for the parameter g. Moreover, when M is large, it is observed that the optimal cells (in the sense of the distortion) are roughly congruent to some basic tessellating cell shape (Gersho's conjecture). Even if it is difficult to find the optimal $\mathfrak{m}(g)$, it could be treated as a constant by admitting Gersho's conjecture since it is normalized.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, Jul. 1948.

[2] S. R. Rajagopalan, L. Sankar, S. Mohajer, and H. V Poor, "Smart meter privacy: A utility-privacy framework", in *Proceedings of IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pp. 190-195, Oct. 2011.

[3] J. Zheng, E. R. Duni and B. D. Rao, "Analysis of Multiple-Antenna Systems With Finite-Rate Feedback Using High-Resolution Quantization Theory", *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1461-1476, April 2007.

[4] M. Kountouris, R. de Francisco, D. Gesbert, D. T. M. Slock and T. Salzer, "Efficient Metrics for Scheduling in MIMO Broadcast Channels with Limited Feedback", in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, 2007, pp. 109-112.

[5] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao and M. Andrews, "An overview of limited feedback in wireless communication systems", *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341-1365, October 2008.

[6] C. K. Au-Yeung, S. Y. Park and D. J. Love, "A Simple Dual-Mode Limited Feedback Multiuser Downlink System", *IEEE Transactions on Communications*, vol. 57, no. 5, pp. 1514-1522, May 2009.

[7] H. Lee, "Comments on "Error Performance of Transmit Beamforming With Delayed and Limited Feedback", *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 585-586, Jan. 2015.

[8] W. Saad, M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems", *IEEE Network*, vol. 34, no. 3, pp. 134-142, May/June 2020.

[9] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies", *IEEE Communications Magazine*, vol. 58, no. 3, pp. 55-61, March 2020.

[10] Emmanuel Bertin; Thomas Magedanz; Noel Crespi, "Toward 6G – Collecting the Research Visions", in *Shaping Future 6G Networks: Needs, Impacts, and Technologies*, IEEE, 2022, pp.1-8.

[11] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. -J. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks", *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84-90, August 2019.

[12] R. M. Gray, and D. L. Neuhoff, "Quantization", *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325-2383,1998.

[13] J. C. Roh and B. D. Rao, "Transmit beamforming in multiple-antenna systems with finite rate feedback: A VQ-based approach", *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1101-1112, March 2006.

[14] S. Rini, L. Barlett, E. Erkip, and Y. C. Eldar, "A general framework for MIMO receivers with low-resolution quantization", in *IEEE Proceedings of Information Theory Workshop (ITW)*, Kaohsiung, Taiwan, Nov. 2017, pp. 599–603.

[15] J. Choi, B. L. Evans, and A. Gatherer, "Resolution-adaptive hybrid MIMO architectures for millimeter wave communications", *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6201–6216, Dec. 2017.

[16] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems", *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[17] N. Shlezinger, Y. C. Eldar and M. R. D. Rodrigues, "Hardware-Limited Task-Based Quantization", *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5223-5238, 2019.

[18] S. Salamtian, N. Shlezinger, Y. C. Eldar, and M. Médard, "Task-based quantization for recovering quadratic functions using principal inertia components", in *Proceedings of IEEE Information Theory Workshop (ITW)*, 2019, pp. 390-394.

[19] K. Choi, K. Tatwawadi, T. Weissman, and S. Ermon, "NECST: neural joint source-channel coding", arXiv preprint arXiv:1811.07557, 2018.

[20] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi, "On Distributed Quantization for Classification", *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 237-249, May 2020.

[21] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli and S. Diggavi, " Quantization of Distributed Data for Learning ", *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 987-1001, Sept. 2021.

[22] F. Sohrabi, K. M. Attiah and W. Yu, "Deep Learning for Distributed Channel Feedback and Multiuser Precoding in FDD Massive MIMO", *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4044-4057, July 2021."

[23] V. Misra, V. K. Goyal and L. R. Varshney, "Distributed Scalar Quantization for Computing: High-Resolution Analysis and Extensions", *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5298-5325, Aug. 2011.

[24] P. Fleischer, "Sufficient conditions for achieving minimum distortion in a quantizer", *IEEE Innt. Conv. Rec.*, pp. 104-111, 1964.

[25] R. Cabral Farias and J. Brossier, "Scalar Quantization for Estimation: From An Asymptotic Design to a Practical Solution", *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2860-2870, June 1, 2014.

[26] M. Kountouris, N. Pappas, "Semantics-empowered communication for networked intelligent systems", *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96-102, 2021.

[27] G. Shi, Y. Xiao, Y. Li and X. Xie, "From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems", *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44-50, August 2021.

[28] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications", *Computer Networks*, vol. 190, 2021.

[29] J. Liu, W. Zhang, H. V. Poor, "A Rate-Distortion Framework for Characterizing Semantic Information", in *Proceedings of IEEE International Symposium on Information Theory (ISIT'21)*, pp. 2894-2899, July 2021.

[30] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems", *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663-2675, 2021.

[31] M. Sana and E. C. Strinati, "Learning semantics: An opportunity for effective 6G communications", in *Proceedings of IEEE 19th Annual Consumer Communications and Networking Conference (CCNC'19)*, 2022.

[32] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission", *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434-2444, 2021.

[33] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things", *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, Jan. 2021.

[34] W. J. Yun, B. Lim, S. Jung, Y.-C. Ko, J. Park, J. Kim, and M. Bennis, "Attention-based reinforcement learning for real-time uav semantic communication", in *Proceedings of 17th IEEE International Symposium on Wireless Communication Systems (ISWCS'17)*, 2021, pp. 1–6.

[35] M. K. Farshbafan, W. Saad, and M. Debbah, "Common Language for Goal-Oriented Semantic Communications: A Curriculum Learning Framework", arXiv preprint arXiv:2111.08051.

[36] Yang, W., Liew, Z.Q., Lim, W.Y.B., Xiong, Z., Niyato, D., Chi, X., Cao, X. and Letaief, K.B., "Semantic Communication Meets Edge Intelligence", arXiv preprint arXiv:2202.06471.

[37] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, " What Is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence", *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 336-371, 2021.

[38] H. Seo, J. Park, M. Bennis, and M. Debbah, "Semantics-Native Communication with Contextual Reasoning", arXiv preprint arXiv:2108.05681.

[39] C. Zhang, N. Khalfet, S. Lasaulce, V. Varma, and S. Tarbouriech, "Payoff-oriented quantization and application to power control", in *Proceeding of 15th IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (Wiopt'17)*, Paris, France, 2017.

[40] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian, and P. Panciatici, "Decision-Oriented Communications: Application to Energy-Efficient Resource Allocation", in *Proceeding of 9th IEEE International Conference on Wireless Networks and Mobile Communications (WINCOM'18)*, Marrakech, Morroco, 2018.

[41] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian and P. Panciatici, "Decision Set Optimization and Energy-Efficient MIMO Communications", in *Proceeding of 30th IEEE International Symposium on Personal,Indoor and Mobile Radio Communications (PIMRC'19)*, Istanbul, Turkey, 2019.

[42] M. R. Garey, D. Johnson, and H. Witsenhausen, "The complexity of the generalized Lloyd-max problem" (corresp.). *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 255-256, March 1982.

[43] P. Samadi, A. H. Mohsenian-Rad, R. Schober R, V.W. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility maximization for smart grid", in *Proceedings of the First IEEE International Conference on Smart Grid Communications*, pp. 415-420, Oct. 2010.

[44] W. Xu, X. Dong, and W. Lu, "MIMO Relaying Broadcast Channels With Linear Precoding and Quantized Channel State Information Feedback", *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5233-5245, Oct. 2010.

[45] B. Makki and T. Eriksson, "Feedback Subsampling in Temporally-Correlated Slowly-Fading Channels using Quantized CSI", *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2282-2294, June 2013.

[46] B. Makki, T. Svensson, T. Eriksson, and M. Debbah, "On Feedback Resource Allocation in Multiple-Input-Single-Output Systems Using Partial CSI Feedback", *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 816-825, March 2015.

[47] E. V. Belmega, and S. Lasaulce, "Energy-Efficient Precoding for Multiple-Antenna Terminals", *IEEE Transactions on Signal Processing*, vol. 59, no. 1, January 2011.

[48] F. Meshkati, A. J. Goldsmith, H. V. Poor, and S. C. Schwartz, "A Game-Theoretic Approach to Energy-Efficient Modulation in CDMA Networks with Delay QoS Constraints", *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 6, pp. 1069-1078.

[49] S. Berri, S. Lasaulce, and M. S. Radjef, "Power control with partial observation in wireless ad hoc networks", in *Proceedings of 24th European Signal Processing Conference (EUSIPCO'16)*, 2016, pp. 1833-1837.

[50] O. Beaude, S. Lasaulce, M. Hennebel, and I. Mohand-Kaci, "Reducing the impact of distributed EV charging on distribution network operating costs", *IEEE Transactions on Smart Grid*, vol. 7, no.6, pp. 2666-2679, June 2016.

[51] O. Motlagh, A. Berry, and L. O'Neil, "Clustering of residential electricity customers using load time series", *Applied Energy*, vol. 237, 2019.

[52] C. Zhang, S. Lasaulce, M. Hennebel, L. Saludjian, P. Panciatici, and H. V. Poor, "Decision-making oriented clustering: Application to pricing and power consumption scheduling", *Applied Energy*, vol. 297, 2021.

[53] A. Gersho, "Asymptotically optimal block quantization", *IEEE Transactions on Information Theory*, vol. 25, pp. 373-380, July 1979.

[54] S. Lloyd, "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, 1982.

[55] J. Max, "Quantizing for minimum distortion", *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7-12, 1960.

[56] Pecan street inc. dataport. [Online]. Available: https://dataport.pecanstreet.org/data.

[57] W. R. Bennett, "Spectra of quantized signal", *The Bell System Technical Journal* , vol. 27, pp. 446-472, July 1948.

[58] P. F. Panter and W. Dite, "Quantizing distortion in pulse-count modulation with nonuniform spacing of levels", *in Proceeding of IRE*, vol. 39, pp. 44-48, Jan. 1951.